~~the probability assignments that maximize entropy. But instead we'll extend the strategy used in the overthinking box on page 274. As a bonus, this strategy will allow us to derive the constraints that are necessary for a distribution, in this case the binomial, to be a maximum entropy distribution.~~

~~Let p be the binomial distribution, and let p_i be the probability of a sequence of observations i with number of successes x_i and number of failures $n - x_i$. Let q be some other discrete distribution defined over the same set of observable sequences. As before, KL divergence tells us that:~~

~~$-H(q,p) \geq H(q) \implies -\sum_i q_i \log p_i \geq -\sum_i q_i \log q_i$~~

~~What we're going to do now is work with $H(q,p)$ and simplify it until we can isolate the constraint~~ that defines the class of distributions for which p has maximum entropy. Let $\lambda = \sum_i p_i x_i$ be the ~~expected value of p. Then from the definition of $H(q,p)$:~~

$$-H(q,p) = -\sum_i q_i \log\left[\left(\frac{\lambda}{n}\right)^{x_i}\left(1-\frac{\lambda}{n}\right)^{n-x_i}\right] = -\sum_i q_i\left(x_i \log\left[\frac{\lambda}{n}\right] + (n-x_i)\log\left[1-\frac{\lambda}{n}\right]\right)$$

After some algebra:

$$-H(q,p) = -\sum_i q_i\left(x_i \log\left[\frac{\lambda}{n-\lambda}\right] + n\log\left[\frac{n-\lambda}{n}\right]\right) = -n\log\left[\frac{n-\lambda}{n}\right] - \log\left[\frac{\lambda}{n-\lambda}\right]\underbrace{\sum_i q_i x_i}_{\bar{q}}$$

~~The term on the far right labeled $\bar{q}$ is the expected value of the distribution q. If we knew it, we could complete the calculation, because no other term depends upon q_i. This means that expected value is the constraint that defines the class of distributions for which the binomial p has maximum entropy. If we now set the expected value of q equal to λ, then $H(q) = H(p)$. For any other expected value of q, $H(p) > H(q)$.~~

~~Finally, notice the term $\log[\lambda/(n-\lambda)]$. This term is the log of the ratio of the expected number of successes to the expected number of failures. That ratio is the "odds" of a success, and its logarithm is called "log odds." This quantity will feature prominently in models we construct from the binomial distribution, in Chapter 11.~~

9.2. Generalized linear models

The Gaussian models of previous chapters worked by first assuming a Gaussian distribution over outcomes. Then, we replaced the parameter that defines the mean of that distribution, μ, with a linear model. This resulted in likelihood definitions of the sort:

$$\begin{aligned} y_i &\sim \text{Normal}(\mu_i, \sigma) \\ \mu_i &= \alpha + \beta x_i \end{aligned}$$

For an outcome variable that is continuous and far from any theoretical maximum or minimum, this sort of Gaussian model has maximum entropy.

But when the outcome variable is either discrete or bounded, a Gaussian likelihood is not the most powerful choice. Consider for example a count outcome, such as the number of blue marbles pulled from a bag. Such a variable is constrained to be zero or a positive integer. Using a Gaussian model with such a variable won't result in a terrifying explosion. But it can't be trusted to do much more than estimate the average count. It certainly can't be trusted to produce sensible predictions, because while you and I know that counts can't be negative, a linear regression model does not. So it would happily predict negative values, whenever the mean count is close to zero.

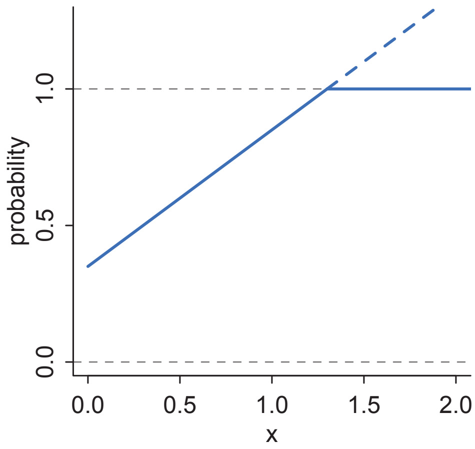

FIGURE 9.5. Why we need link functions. The solid blue line is a linear model of a probability mass. It increases linearly with a predictor, x, on the horizontal axis. But when it reaches the maximum probability mass of 1, at the dashed boundary, it will happily continue upwards, as shown by the dashed blue line. In reality, further increases in x could not further increase probability, as indicated by the horizontal continuation of the solid trend.

Luckily, it's easy to do better. By using all of our prior knowledge about the outcome variable, usually in the form of constraints on the possible values it can take, we can appeal to maximum entropy for the choice of distribution. Then all we have to do is generalize the linear regression strategy—replace a parameter describing the shape of the likelihood with a linear model—to probability distributions other than the Gaussian.

This is the essence of a **GENERALIZED LINEAR MODEL**.[136] And it results in models that look like this:

$$
\begin{aligned}
y_i &\sim \text{Binomial}(n, p_i) \\
f(p_i) &= \alpha + \beta x_i
\end{aligned}
$$

There are only two changes here from the familiar Gaussian model. The first is principled—the principle of maximum entropy. The second is an epicycle—a modeling trick that works descriptively but not causally—but a quite successful one. I'll briefly explain each, before moving on in the remainder of the section to describe all of the most common distributions used to construct generalized linear models. Later chapters show you how to implement them.

First, the likelihood is binomial instead of Gaussian. For a count outcome y for which each observation arises from n trials and with constant expected value np, the binomial distribution has maximum entropy. So it's the least informative distribution that satisfies our prior knowledge of the outcomes y. If the outcome variable had different constraints, it could be a different maximum entropy distribution.

Second, there is now a funny little f at the start of the second line of the model. This represents a **LINK FUNCTION**, to be determined separately from the choice of distribution. Generalized linear models need a link function, because rarely is there a "μ", a parameter describing the average outcome, and rarely are parameters unbounded in both directions, like μ is. For example, the shape of the binomial distribution is determined, like the Gaussian, by two parameters. But unlike the Gaussian, neither of these parameters is the mean. Instead, the mean outcome is np, which is a function of both parameters. Since n is usually known (but not always), it is most common to attach a linear model to the unknown part, p. But p is a probability mass, so p_i must lie between zero and one. But there's nothing to stop the linear model $\alpha + \beta x_i$ from falling below zero or exceeding one. FIGURE 9.5 plots an example.

The link function *f* provides a solution to this common problem. This chapter will introduce the two most common link functions. Then you'll see how to use them in the chapters that follow.

Rethinking: The scourge of Histomancy. One strategy for choosing an outcome distribution is to plot the histogram of the outcome variable and, by gazing into its soul, decide what sort of distribution function to use. Call this strategy HISTOMANCY, the ancient art of divining likelihood functions from empirical histograms. This sorcery is used, for example, when testing for normality before deciding whether or not to use a non-parametric procedure. Histomancy is a false god, because even perfectly good Gaussian variables may not look Gaussian when displayed as a histogram. Why? Because at most what a Gaussian likelihood assumes is not that the aggregated data look Gaussian, but rather that the *residuals*, after fitting the model, look Gaussian. So for example the combined histogram of male and female body weights is certainly not Gaussian. But it is (approximately) a mixture of Gaussian distributions, so after conditioning on sex, the residuals may be quite normal. Other times, people decide not to use a Poisson model, because the variance of the aggregate outcome exceeds its mean (see Chapter 10). But again, at most what a Poisson likelihood assumes is that the variance equals the mean after conditioning on predictors. It may very well be that a Gaussian or Poisson likelihood is a poor assumption in any particular context. But this can't easily be decided via Histomancy. This is why we need principles, whether maximum entropy or otherwise.

9.2.1. Meet the family. The most common distributions used in statistical modeling are members of a family known as the EXPONENTIAL FAMILY. Every member of this family is a maximum entropy distribution, for some set of constraints. And conveniently, just about every other statistical modeling tradition employs the exact same distributions, even though they arrive at them via justifications other than maximum entropy.

FIGURE 9.6 illustrates the representative shapes of the most common exponential family distributions used in GLMs. The horizontal axis in each plot represents values of a variable, and the vertical axis represents probability density (for the continuous distributions) or probability mass (for the discrete distributions). For each distribution, the figure also provides the notation (above each density plot) and the name of R's corresponding built-in distribution function (below each density plot). The gray arrows in FIGURE 9.6 indicate some of the ways that these distributions are dynamically related to one another. These relationships arise from generative processes that can convert one distribution to another. You do not need to know these relationships in order to successfully use these distributions in your modeling. But the generative relationships do help to demystify these distributions, by tying them to causation and measurement.

Two of these distributions, the Gaussian and binomial, are already familiar to you. Together, they comprise the most commonly used outcome distributions in applied statistics, through the procedures of linear regression (Chapter 4) and logistic regression (Chapter 10). There are also three new distributions that deserve some commentary.

The EXPONENTIAL DISTRIBUTION (center) is constrained to be zero or positive. It is a fundamental distribution of distance and duration, kinds of measurements that represent displacement from some point of reference, either in time or space. If the probability of an event is constant in time or across space, then the distribution of events tends towards exponential. The exponential distribution has maximum entropy among all non-negative continuous distributions with the same average displacement. Its shape is described by a single parameter, the rate of events λ, or the average displacement λ^{-1}. This distribution is the core of survival and event history analysis, which is not covered in this book.

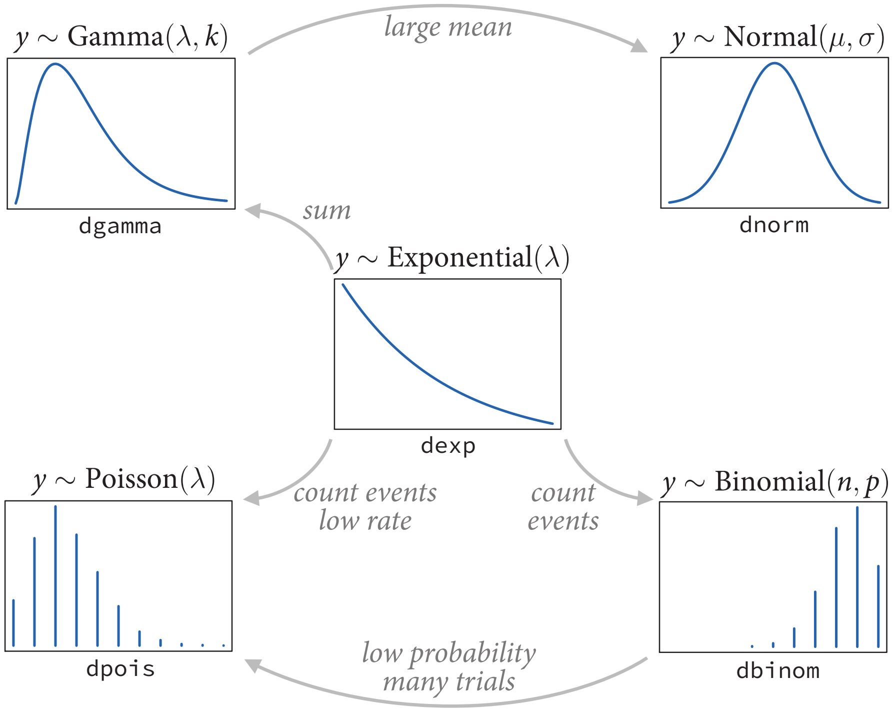
$y \sim \text{Gamma}(\lambda, k)$
dgamma
large mean
$y \sim \text{Normal}(\mu, \sigma)$
dnorm
sum
$y \sim \text{Exponential}(\lambda)$
dexp
$y \sim \text{Poisson}(\lambda)$
dpois
count events
low rate
count
events
$y \sim \text{Binomial}(n, p)$
dbinom
low probability
many trials

FIGURE 9.6. Some of the exponential family distributions, their notation, and some of their relationships. Center: exponential distribution. Clockwise, from top-left: gamma, normal (Gaussian), binomial and Poisson distributions.

The **GAMMA DISTRIBUTION** (top-left) is also constrained to be zero or positive. It too is a fundamental distribution of distance and duration. But unlike the exponential distribution, the gamma distribution can have a peak above zero. If an event can only happen after two or more exponentially distributed events happen, the resulting waiting times will be gamma distributed. For example, age of cancer onset is approximately gamma distributed, since multiple events are necessary for onset.[137] The gamma distribution has maximum entropy among all distributions with the same mean and same average logarithm. Its shape is described by two parameters, but there are at least three different common descriptions of these parameters, so some care is required when working with it. The gamma distribution is common in survival and event history analysis, as well as some contexts in which a continuous measurement is constrained to be positive.

The **POISSON DISTRIBUTION** (bottom-left) is a count distribution like the binomial. It is actually a special case of the binomial, mathematically. If the number of trials n is very large (and usually unknown) and the probability of a success p is very small, then a binomial distribution converges to a Poisson distribution with an expected rate of events per unit time of $\lambda = np$. Practically, the Poisson distribution is used for counts that never get close to any theoretical maximum. As a special case of the binomial, it has maximum entropy under exactly the same constraints. Its shape is described by a single parameter, the rate of events λ. Poisson GLMs are detailed in the next chapter.

There are many other exponential family distributions, and many of them are useful. But don't worry that you need to memorize them all. You can pick up new distributions, and the sorts of generative processes they correspond to, as needed. It's also not important that an outcome distribution be a member of the exponential family—if you think you have good reasons to use some other distribution, then use it. But you should also check its performance, just like you would any modeling assumption.

~~**Rethinking: A likelihood is a prior.** In traditional statistics, likelihood functions are "objective" and prior distributions "subjective." However, likelihoods are themselves prior probability distributions: They are priors for the data, conditional on the parameters. And just like with other priors, there is no correct likelihood. But there are better and worse likelihoods, depending upon the context. Useful inference does not require that the data (or residuals) be actually distributed according to the likelihood anymore than it requires the posterior distribution to be like the prior.~~

9.2.2. Linking linear models to distributions. To build a regression model from any of the exponential family distributions is just a matter of attaching one or more linear models to one or more of the parameters that describe the distribution's shape. But as hinted at earlier, usually we require a **LINK FUNCTION** to prevent mathematical accidents like negative distances or probability masses that exceed 1. So for any outcome distribution, say for example the exotic "Zaphod" distribution,[138] we write:

$$\begin{aligned} y_i &\sim \text{Zaphod}(\theta_i, \phi) \\ f(\theta_i) &= \alpha + \beta x_i \end{aligned}$$

where *f* is a link function.

But what function should *f* be? A link function's job is to map the linear space of a model like $\alpha + \beta x_i$ onto the non-linear space of a parameter like θ. So *f* is chosen with that goal in mind. Most of the time, for most GLMs, you can use one of two exceedingly common links, a *logit link* or a *log link*. Let's introduce each, and you'll work with both in later chapters.

The **LOGIT LINK** maps a parameter that is defined as a probability mass, and therefore constrained to lie between zero and one, onto a linear model that can take on any real value. This link is extremely common when working with binomial GLMs. In the context of a model definition, it looks like this:

$$\begin{aligned} y_i &\sim \text{Binomial}(n, p_i) \\ \text{logit}(p_i) &= \alpha + \beta x_i \end{aligned}$$

And the logit function itself is defined as the *log-odds*:

$$\text{logit}(p_i) = \log \frac{p_i}{1 - p_i}$$

The "odds" of an event are just the probability it happens divided by the probability it does not happen. So really all that is being stated here is:

$$\log \frac{p_i}{1 - p_i} = \alpha + \beta x_i$$

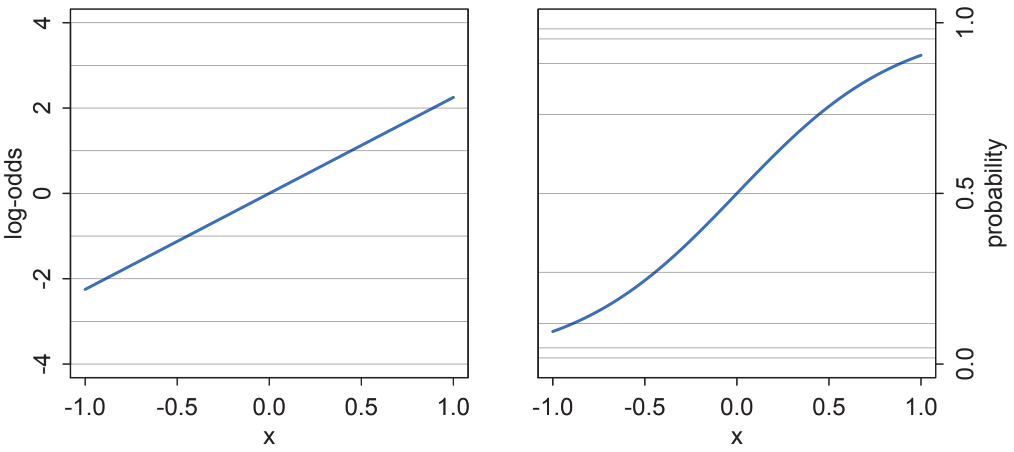

FIGURE 9.7. The logit link transforms a linear model (left) into a probability (right). This transformation compresses the geometry far from zero, such that a unit change on the linear scale (left) means less and less change on the probability scale (right).

So to figure out the definition of p_i implied here, just do a little algebra and solve the above equation for p_i:

$$p_i = \frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)}$$

The above function is usually called the **LOGISTIC**. In this context, it is also commonly called the **INVERSE-LOGIT**, because it inverts the logit transform.

What all of this means is that when you use a logit link for a parameter, you are defining the parameter's value to be the logistic transform of the linear model. FIGURE 9.7 illustrates the transformation that takes place when using a logit link. On the left, the geometry of the linear model is shown, with horizontal lines indicating unit changes in the value of the linear model as the value of a predictor x changes. This is the log-odds space, which extends continuously in both positive and negative directions. On the right, the linear space is transformed and is now constrained entirely between zero and one. The horizontal lines have been compressed near the boundaries, in order to make the linear space fit within the probability space. This compression produces the characteristic logistic shape of the transformed linear model shown in the right-hand plot.

This compression does affect interpretation of parameter estimates, because no longer does a unit change in a predictor variable produce a constant change in the mean of the outcome variable. Instead, a unit change in x_i may produce a larger or smaller change in the probability p_i, depending upon how far from zero the log-odds are. For example, in FIGURE 9.7, when $x = 0$ the linear model has a value of zero on the log-odds scale. A half-unit increase in x results in about a 0.25 increase in probability. But each addition half-unit will produce less and less of an increase in probability, until any increase is vanishingly small. And if you think about it, a good model of probability needs to behave this way. When an

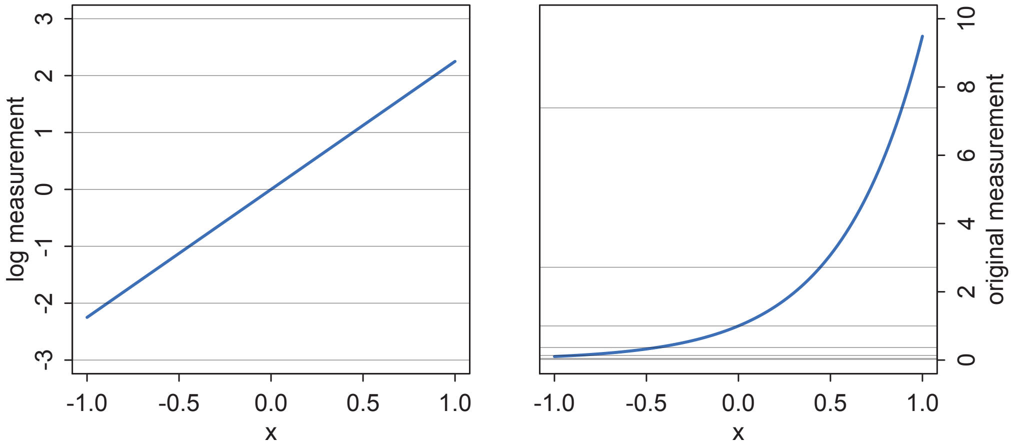

FIGURE 9.8. The log link transforms a linear model (left) into a strictly positive measurement (right). This transform results in an exponential scaling of the linear model, with a unit change on the linear scale mapping onto increasingly larger changes on the outcome scale.

event is almost guaranteed to happen, its probability cannot increase very much, no matter how important the predictor may be.

You'll find examples of this compression phenomenon in later chapters. The key lesson for now is just that no regression coefficient, such as β, from a GLM ever produces a constant change on the outcome scale. Recall that we defined interaction (Chapter 7) as a situation in which the effect of a predictor depends upon the value of another predictor. Well now every predictor essentially interacts with itself, because the impact of a change in a predictor depends upon the value of the predictor before the change. More generally, every predictor variable effectively interacts with every other predictor variable, whether you explicitly model them as interactions or not. This fact makes the visualization of counter-factual predictions even more important for understanding what the model is telling you.

The second very common link function is the **LOG LINK**. This link function maps a parameter that is defined over only positive real values onto a linear model. For example, suppose we want to model the standard deviation σ of a Gaussian distribution so it is a function of a predictor variable x. The parameter σ must be positive, because a standard deviation cannot be negative nor can it be zero. The model might look like:

$$y_i \sim \text{Normal}(\mu, \sigma_i)$$
$$\log(\sigma_i) = \alpha + \beta x_i$$

In this model, the mean μ is constant, but the standard deviation scales with the value x_i. A log link is both conventional and useful in this situation. It prevents σ from taking on a negative value.

What the log link effectively assumes is that the parameter's value is the exponentiation of the linear model. Solving $\log(\sigma_i) = \alpha + \beta x_i$ for σ_i yields the inverse link:

$$\sigma_i = \exp(\alpha + \beta x_i)$$

The impact of this assumption can be seen in FIGURE 9.8. Using a log link for a linear model (left) implies an exponential scaling of the outcome with the predictor variable (right). Another way to think of this relationship is to remember that logarithms are *magnitudes*. An increase of one unit on the log scale means an increase of an order of magnitude on the untransformed scale. And this fact is reflected in the widening intervals between the horizontal lines in the right-hand plot of FIGURE 9.8.

While using a log link does solve the problem of constraining the parameter to be positive, it may also create a problem when the model is asked to predict well outside the range of data used to fit it. Exponential relationships grow, well, exponentially. Just like a linear model cannot be linear forever, an exponential model cannot be exponential forever. Human height cannot be linearly related to weight forever, because very heavy people stop getting taller and start getting wider. Likewise, the property damage caused by a hurricane may be approximately exponentially related to wind speed for smaller storms. But for very big storms, damage may be capped by the fact that everything gets destroyed.

Rethinking: When in doubt, play with assumptions. Link functions do amount to assumptions. And like all assumptions, they are useful in different contexts. The conventional logit and log links are widely useful, but they can sometimes distort inference. If you ever have doubts, and want to reassure yourself that your conclusions are not sensitive to choice of link function, then do what you'd do for any other modeling assumption: SENSITIVITY ANALYSIS. A sensitivity analysis explores how changes in assumptions influence inference. If none of the alternative assumptions you consider have much impact on inference, that's worth reporting. Likewise, if the alternatives you consider do have an important impact on inference, that's also worth reporting. The same sort of advice follows for other modeling assumptions: likelihoods, linear models, priors, and even how the model is fit to data. As with many machines, exploring how a model behaves under extreme conditions helps us understand how it behaves under ordinary conditions.

Some people are nervous about sensitivity analysis, because it feels like fishing for results, or "*p*-hacking."[139] The goal of sensitivity analysis is really the opposite of *p*-hacking. In *p*-hacking, many justifiable analyses are tried, and the one that attains statistical significance is reported. In sensitivity analysis, many justifiable analyses are tried, and all of them are described.

Overthinking: Parameters interacting with themselves. We can find some further clarity on the claim that GLMs force every predictor variable to interact with itself by mathematically computing the rate of change in the outcome for a given change in the value of the predictor. First, recall that in a classic Gaussian model the mean is modeled like:

$$\mu = \alpha + \beta x$$

So the rate of change in μ with respect to x is just $\partial\mu/\partial x = \beta$. And that's constant. It doesn't matter what value x has. But now consider the rate of change in a binomial probability p with respect to a predictor x:

$$p = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}$$

And now taking the derivative with respect to x yields:

$$\frac{\partial p}{\partial x} = \frac{\beta}{2\big(1 + \cosh(\alpha + \beta x)\big)}$$

Since x appears in this answer, the impact of a change in x depends upon x. That's an interaction with itself.

9.2.3. Absolute and relative differences. There is an important practical consequence of the way that a link function compresses and expands different portions of the linear model's range: Parameter estimates do not by themselves tell you the importance of a predictor on the outcome. The reason is that each parameter represents a *relative* difference on the scale of the linear model, ignoring other parameters, while we are really interested in *absolute* differences in outcomes that must incorporate all parameters.

This point will come up again in the context of data examples in later chapters, when it will be easier to illustrate its importance. For now, just keep in mind that a big beta-coefficient may not correspond to a big effect on the outcome.

~~**9.2.4. GLMs and information criteria.** What you learned in Chapter 6 about information criteria and regularizing priors applies also to GLMs. But with all these new outcome distributions at your command, it is tempting to use information criteria to compare models with different likelihood functions. Is a Gaussian or binomial better? Can't we just let WAIC sort it out?~~

~~Unfortunately, WAIC (or any other information criterion) cannot sort it out. The problem is that deviance is part normalizing constant. The constant affects the absolute magnitude of the deviance, but it doesn't affect fit to data. Since information criteria are all based on deviance, their magnitude also depends upon these constants. That is fine, as long as all of the models you compare use the same outcome distribution type—Gaussian, binomial, exponential, gamma, Poisson, or another. In that case, the constants subtract out when you compare models by their differences. But if two models have different outcome distributions, the constants don't subtract out, and you can be misled by a difference in AIC/DIC/WAIC.~~

~~Really all you have to remember is to only compare models that all use the same type of likelihood. Of course it is possible to compare models that use different likelihoods, just not with information criteria. Luckily, the principle of maximum entropy ordinarily motivates an easy choice of likelihood, at least for ordinary regression models. So there is no need to lean on information criteria for this modeling choice.~~

~~There are a few nuances with WAIC and individual GLM types. These nuances will arise as examples of each GLM are worked, in later chapters.~~

~~9.3. Maximum entropy priors~~

~~The principle of maximum entropy helps us to make modeling choices. When pressed to choose an outcome distribution—a likelihood—maximum entropy nominates the least informative distribution consistent with the constraints on the outcome variable. Applying the principle in this way leads to many of the same distributional choices that are commonly regarded as just convenient assumptions or useful conventions.~~

~~Another way that the principle of maximum entropy helps with choosing distributions arises when choosing priors. GLMs are easy to use with conventional weakly informative priors of the sort you've been using up to this point in the book. Such priors are nice, because they allow the data to dominate inference while also taming some of the pathologies of unconstrained estimation. There were some striking examples of their "soft power" in Chapter 8.~~

~~But sometimes, rarely, some of the parameters in a GLM refer to things we might actually have background information about. When that's true, maximum entropy provides a way to generate a prior that embodies the background information, while assuming as little else as possible. This makes them appealing, conservative choices.~~

~~We won't be using maximum entropy to choose priors in this book, but when you come across an analysis that does, you can interpret the principle in the same way as you do with likelihoods and understand the approach as an attempt to include relevant background information about parameters, while introducing no other assumptions by accident.~~

9.4. Summary

This chapter has been a conceptual, not practical, introduction to maximum entropy and generalized linear models. The principle of maximum entropy provides an empirically successful way to choose likelihood functions. Information entropy is essentially a measure of the number of ways a distribution can arise, according to stated assumptions. By choosing the distribution with the biggest information entropy, we thereby choose a distribution that obeys the constraints on outcome variables, without importing additional assumptions. Generalized linear models arise naturally from this approach, as extensions of the linear models in previous chapters. The necessity of choosing a link function to bind the linear model to the generalized outcome introduces new complexities in model specification, estimation, and interpretation. You'll become comfortable with these complexities through examples in later chapters.